

ABSTRAK

Web crawler adalah salah satu bagian penting dari *search engine*. *Web crawler* bertugas mengambil informasi dari halaman web yang dikunjungi. Untuk mendapatkan konten khusus bahasa Indonesia telah didesain dan diimplementasikan pada penelitian sebelumnya oleh Arifin(2011). Dari hasil penelitian ini diperoleh *web crawler* yang mampu menghimpun *website* bahasa Indonesia dengan data ujicoba 1000 *url*, 500 *website* bahasa Indonesia dan 500 *website* asing. Namun data yang tersimpan dari *web crawler* tersebut belum dapat optimal, karena di dalam konten yang terambil masih terdapat beberapa bagian yang bukan sebuah informasi.

Kekurangan pada pengambilan konten tersebut dicoba untuk diperbaiki dalam penelitian ini dengan menggunakan ekstraksi informasi(*web data extraction*) berbasis pencocokan pohon dengan menggunakan metode *simple tree matching* (STM). Ekstraksi informasi adalah sebuah cara computer dalam pengambilan informasi dalam suatu *website*. *Simple tree matching* adalah sebuah metode ekstraksi informasi yang melakukan pencocokan dua *tree website* pada proses ekstraksinya.

Dengan metode *simple tree matching*, *web crawler* RISE dilakukan ujicoba kembali pada 50 *website* dari data penelitian sebelumnya dan didapatkan hasil konten yang lebih bersih dari implementasi sebelumnya. Dapat disimpulkan bahwa *simple tree matching* mampu dengan sempurna melakukan ekstraksi informasi pada *web crawler*. Dengan rata-rata nilai *precision* 100% dan *recall* 89.31365% .

Kata Kunci : *Search engine, web crawler, web data extraction, simple tree matching*

(Halaman ini sengaja dikosongkan)

KATA PENGANTAR

Segala puji bagi Allah *Subhanahuwata'ala* karena atas rahmat dan hidayah-Nya sehingga laporan Tugas Akhir dengan judul “Penerapan Ekstraksi Informasi *Simple Tree Matching*(STM) pada *web crawler* Bahasa Indonesia” dapat terselesaikan dengan baik dan tepat pada waktunya.

Penulis menyadari bahwa masih banyak kekurangan dan ketidaksempurnaan pada penulisan maupun pemrograman yang telah dibuat, namun dari ketidaksempurnaan inilah yang akan mendorong seseorang menuju arah yang lebih baik guna penyempurnaannya.

Selama penulisan Tugas Akhir ini, Penulis menyadari banyak pihak yang telah memberikan waktunya untuk membantu dalam penyelesaian Tugas Akhir ini untuk semua itu Penulis ingin mengucapkan terima kasih yang tulus kepada:

1. Bapak Firdaus Solihin, S.Kom., M.Kom. selaku ketua jurusan teknik informatika sekaligus dosen pembimbing I
2. Bapak Iwan Santosa, ST., MT. selaku koordinator bidang minat sistem terdistribusi sekaligus dosen pembimbing II
3. Ibu Fika Hastarita, S.T., M.Eng. selaku kordinator tugas akhir
4. Seluruh Staf pengajar Teknik Informatika atas ilmu yang diberikan selama proses perkuliahan, semoga Allah menjadikan semua ini amal sholeh yang bermanfaat.
5. Staf Tata Usaha Jurusan atas bantuan melayani segala administrasi selama perkuliahan ini.
6. Seluruh anggota keluarga untuk doa dan dukungannya selama ini yang telah diberikan tanpa pamrih.
7. Bapak Encik Muhammad Fauzan, S.H., LLM. dan seluruh staff Satuan Penjaminan Mutu yang menyediakan kesempatan berkarir dan waktu untuk menyelesaikan tugas akhir ini.
8. Teman-teman seperjuangan teknik informatika angkatan 2009(Solite'09) di universitas trunojoyo madura.
9. Spesial mas M. Syaiful Arifin, S.Kom. yang sempat-secukupnya meluangkan waktunya untuk membantu dalam hal apapun.
10. Semua pihak yang tidak dapat kami sebutkan satu persatu, yang telah membantu dalam penyusunan laporan Tugas Akhir.

Penulis menyadari bahwa dalam pembuatan tugas akhir ini masih terdapat kekurangan-kekurangan. Untuk itu saran dan kritik membangun

sangat penulis harapkan demi kesempurnaan tugas akhir ini. Akhir kata, penulis berharap tugas akhir ini dapat bermanfaat bagi semua pihak. Amien Yaa Robbaal Alamiin.

Bangkalan, 05 Juni 2013

Penulis

DAFTAR ISI

Halaman pengesahan.....	iii
Persembahan	vii
Abstrak.....	ix
Kata pengantar	xi
Daftar isi.....	xiii
Daftar gambar	xv
Daftar tabel.....	xvii
Daftar istilah.....	xix
Daftar lampiran	xxi
Bab I.....	1
1.1.Latar belakang.....	1
1.2.Perumusan masalah.....	2
1.3.Tujuan dan manfaat.....	2
1.4.Batasan masalah.....	2
1.5.Metodologi penelitian	3
1.5.1.Studi literatur.....	3
1.5.2.Studi lapangan.....	3
1.5.3.Analisa dan perancangan sistem	4
1.5.4.Pembuatan dan pengujian sistem	4
1.5.5.Penarikan kesimpulan	4
1.6.Sistematika	4
Bab II	7
2.1.Tinjauan pustaka	7
2.1.1. Desain dan implementasi web crawler penghimpun website bahasa indonesia	7
2.1.2. Language based crawling: crawling the arabic content of the web.....	9
2.1.3. Information extraction from web pages using automatic pattern discovery method based on tree matching	10
2.1.4. Web data extraction based on simple tree matching	12
2.2. Landasan teori	13
2.2.1. Sistem temu kembali informasi	13
2.2.2. Mesin pencarian	15
2.2.3. Web crawler	16
2.2.4. Ekstraksi informasi	17
2.2.5. Membangun html tag tree	18

2.2.6. Mining data regions	19
2.2.7. Mining data records	20
2.2.8. Tree edit distance	23
2.2.9. Simple tree matching	24
2.2.10. Partial tree alignment	28
2.2.11. Regular expression.....	31
2.2.12. Stopword.....	32
2.2.13. Url and text normalizer	35
Bab III.....	36
3.1. Web crawler.....	37
3.1.1. Online module	40
3.1.2. Offline module.....	43
3.2. Perancangan basis data	50
Bab IV	53
5.1. Batasan implementasi	53
5.2. Lingkungan implementasi.....	53
4.3. Pembuatan program	54
4.5. Implementasi program	64
4.7. Evaluasi	70
Bab V	80
5.1. Kesimpulan.....	81
5.2. Saran	81
Daftar pustaka	82
Lampiran-lampiran	85
Daftar riwayat hidup	103

DAFTAR GAMBAR

Gambar 1.1. Alur metodologi penelitian.....	3
Gambar 2.1. Desain Web Crawler bahasa indonesia	8
Gambar 2.2. Arsitektur <i>web crawler</i> berbasis bahasa	9
Gambar 2.3. Proses <i>DEPTA</i>	11
Gambar 2.4. <i>Pseudocode Simple Tree Matching</i>	13
Gambar 2.5. Arsitektur dasar <i>Information Retrieval System</i>	14
Gambar 2.6. Arsitektur <i>High level Google Search Engine</i>	15
Gambar 2.7. <i>Flowchart</i> proses dasar sistem crawler.....	17
Gambar 2.8. Bagan Ekstraksi Informasi	18
Gambar 2.9. Halaman web yang menjelaskan data record dan data region.....	21
Gambar 2.10. Halaman web dengan dua <i>data record</i>	22
Gambar 2.11. Contoh tag tree HTML dari halaman web pada Gambar 4	23
Gambar 2.12. <i>Pseudocode Simple Tree Matching</i>	25
Gambar 2.13. Ilustrasi <i>Simple Tree Matching</i>	26
Gambar 2.14. <i>Web data extraction base on partial tree alignment</i>	28
Gambar 2.15. Memperluas seed tree.....	30
Gambar 2.16. Ilustrasi <i>Partial Tree Alignment</i>	30
Gambar 3.1. <i>Flowchart web crawler</i>	37
Gambar 3.2. <i>Flowchart main scheduler</i>	39
Gambar 3.3. <i>Flowchart Configure Loader</i>	40
Gambar 3.4. <i>Flowchart Proses URL Inserter</i>	41
Gambar 3.5. <i>Flowchart online scheduler</i>	41
Gambar 3.6. <i>Flowchart multithread downloader</i>	42
Gambar 3.7. <i>Flowchart Offline Scheduler</i>	44
Gambar 3.8. <i>Flowchart language detector</i>	45
Gambar 3.9. <i>Flowchart ekstraksi link</i>	46
Gambar 3.10. Alur <i>Structured Data Extractions</i>	47
Gambar 3.11. Contoh halaman web yang akan di ekstraksi.....	48
Gambar 3.12. Hasil Ekstraksi <i>data record</i> berupa tabel setiap <i>region</i>	48
Gambar 3.13. Data items dibandingkan dengan template sejajar.....	49
Gambar 3.14. Penghapusan tabel setelah dibandingkan	49
Gambar 3.15. CDM Web Crawler RISE Project.....	51

Gambar 3.16. PDM <i>Web Crawler RISE project</i>	51
Gambar 4.1. Perangkat Keras untuk uji coba/implementasi	54
Gambar 4.2. File Konfigurasi RiseCrawler	66
Gambar 4.3. Insert data url ke <i>database</i>	67
Gambar 4.4. File <i>URL inserter</i>	67
Gambar 4.5. <i>Running Program RISE WebCrawler</i>	68
Gambar 4.6. Proses <i>Download</i> pada <i>RISE Crawler</i>	68
Gambar 4.7. Proses <i>Link Ekstraktor</i> pada <i>RISE Crawler</i>	68
Gambar 4.8. Proses <i>SDE</i> pada <i>RISE Crawler</i>	69
Gambar 4.9. Folder Penyimpanan pada <i>RISE Crawler</i>	69
Gambar 4.10. File yang tersimpan pada <i>RISE Crawler</i>	69
Gambar 4.11. Grafik perbandingan kapasistas penyimpanan <i>RISE Crawler</i>	71
Gambar 4.12. Grafik perbandingan record tersimpan pada <i>RISE Crawler</i>	72
Gambar 4.13. <i>Error Web Crawler</i> Penelitian Sebelumnya.....	72

DAFTAR TABEL

Tabel 2.1. Notasi <i>Regular Expression</i>	31
Tabel 2.2. Stoplist dari kata dasar stopword bahasa indonesia	33
Tabel 4.1. Perbandingan kapasitas penyimpanan RISE Crawler	70
Tabel 4.2. Perbandingan <i>record</i> tersimpan pada RISE Crawler.....	71
Tabel 4.3. <i>Precision</i> dan <i>Recall</i> dari data aktual dan hasil <i>crawler</i> ...	73

(Halaman ini sengaja dikosongkan)

DAFTAR ISTILAH

Crawler	: Aplikasi yang berfungsi untuk mengambil semua informasi yang diperlukan dengan cara menjelajahi link-link yang ada di halaman situs web kemudian disimpan pada database
Data raw	: Data murni yang belum di olah
Data Record	: Data terstruktur dalam halaman web
Data Region	: Sekumpulan <i>data record</i> yang mempunyai definisi yang sama
DEPTA	: <i>Data Ekstraction based on Partial Tree Alignmen</i>
Font	: Jenis-jenis huruf yang digunakan dalam suatu tulisan teks komputer
Indexing	: Proses pemetaan kata
RISE Project	: Research for Indonesian Search Engine
Parsing	: Proses pembentukan Pohon Sintaks untuk suatu kalimat
Partial Tree Alignment	: Algoritma penjajaran <i>data item</i> pada <i>data record</i>
Protocol	: Aturan-aturan atau tatacara yang digunakan dalam melaksanakan pertukaran data dalam sebuah jaringan
Proxy	: Sebuah fungsi/tugas yang dijalankan suatu Perangkat jaringan, biasanya sebagai perantara antara client dan server
Query	: Permintaan informasi dari database
Regular expressions	: Notasi yang digunakan untuk mendeskripsikan himpunan karakter string
SDE	: <i>Structured data extraction</i>
Search engine	: Alat atau aplikasi yang berfungsi untuk mencari informasi berdasarkan kata kunci
STM	: <i>Simple Tree Matching</i> / metode pencocokan

berbasis pohon untuk mengekstrak data terstruktur

<i>Stopword</i>	: Kata yang banyak digunakan tapi kurang memiliki arti
<i>Tag html</i>	: Sebuah kode yang digunakan dalam HTML untuk menentukan perubahan format
<i>Template</i>	: Tata letak sebagai acuan penempatan sesuatu
<i>Thread</i>	: Sebuah alur kontrol dari sebuah proses
<i>Tokenizer</i>	: TokenStream dan bertanggung jawab untuk memecah text yang masuk ke dalam token
<i>User interface</i>	: Cara program dan user berkomunikasi
<i>Web Content</i>	: Isi dari halaman web

DAFTAR LAMPIRAN

Lampiran 1. <i>Stoplist</i> menggunakan <i>regular expression</i>	82
Lampiran 2. Kode Status <i>HTML Response</i>	92
Lampiran 3. Data Uji Coba 50 <i>website</i> dari penelitian sebelumnya ...	94